RAND

Assessing Research

The Researchers' View

Steven Wooding, Jonathan Grant

RAND Europe

RAND

Assessing Research

The Researchers' View

Steven Wooding, Jonathan Grant

Prepared for the
Higher Education Funding Council for England

RAND Europe

The research described in this report was prepared for the Higher Education Funding Council for England.

ISBN: 0-8330-3480-4

Published 2003 by the RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Contents

Cover Image: Part of the analysis of the research assessment systems designed by the workshop participants.

Preface

This report, prepared for and funded by the Joint Funding Bodies' Review of Research Assessment, presents findings from a series of nine facilitated workshops held with academics and research managers across the UK in December 2002. The objective of the workshops was to investigate views of research quality and attitudes towards different models of research assessment. This report is intended to inform the deliberations of Sir Gareth Roberts' review[1] of how research at UK higher education institutions is assessed.

The report outlines the recurring themes and issues raised by the 142 participants in the workshops. The participants, academics and research managers, represented over one third of the 173 institutions that submitted to the Research Assessment Exercise in 2001.

This report will be of interest to those concerned with research assessment and evaluation in academic research, both practitioners and policy makers. The report provides a survey of the issues of importance and concern to those involved in academic research and the management of that research. The first volume of this report describes the methodology and details the findings of the work. The second volume contains additional source data. Both of the publications can be obtained from the website of the Higher Education Funding Council for England (http://www.hefce.ac.uk/research) or RAND Europe (http://www.randeurope.org, publication number: MR-1698-HEFCE).

RAND Europe is an independent not-for-profit policy research organization that serves the public interest by improving policymaking and informing public debate. Our clients are European governments, institutions, and firms with a need for rigorous, impartial, multidisciplinary analysis of the hardest problems they face. This report has been peer-reviewed in accordance with RAND's quality assurance standards (see http://www.rand.org/about/standards/) and therefore may be represented as a RAND Europe product.

For more information about RAND Europe or this project, please contact:

Martijn van der Mandele, President
RAND Europe
Newtonweg 1
2333 CP Leiden
The Netherlands
Tel: +31 71 524 5151
Fax: +31 71 524 5191
email: reinfo@rand.org

Jonathan Grant, Research Leader
RAND Europe
Grafton House
64 Maids Causeway
Cambridge
CB5 8DD
Tel: +44 1223 353329
Tel: +44 1223 358845

[1] http://www.ra-review.ac.uk/

Acknowledgements

Vanessa Conte and Sian Thomas at HEFCE provided invaluable advice and assistance with this project. We would also like to thank all the participants who attended the workshops, and the administrators who organised the workshops and recruited the attendees.

At RAND Europe we would like to thank Rebecca Shoob who handled our travel arrangements with accuracy and aplomb; and Suja Sivadasan who reviewed and commented on the final report. Finally, we are grateful to the staff of the RAND Europe Cambridge office who participated in the pilot workshop and also assisted with the data analysis.

Executive Summary

Introduction

This Executive Summary presents the key findings from a series of nine facilitated workshops with academics and research managers, which examined how research in UK Higher Education Institutions (HEIs) could be assessed.

Every five years the Research Assessment Exercise (RAE) evaluates the quality of research in UK HEIs. The RAE results are used to allocate resources in a way that rewards excellence. After the most recent exercise in 2001 (RAE2001) the UK Higher Education Funding Bodies were unable to provide the estimates £170m of extra funding required to reward the increase in research excellence revealed by the exercise. This led the funding bodies to start a review of the system used for research assessment, of which this report forms a part.

The full report, 'Assessing Research: The Researchers' View', is split into two volumes and can be obtained from the websites of both The Higher Education Funding Council for England (www.hefce.ac.uk) and RAND Europe (www.randeurope.org).

Methodology

Facilitated workshops were used to stimulate broad and innovative thinking about research assessment, while at the same time providing a structure that would allow comparison between workshop groups. As with all qualitative research methodologies the facilitated workshop provides a method of gaining an insight into attitudes and opinions; however, this insight is – by its very nature – not necessarily representative of the population sampled in a statistical sense. It is also possible, that even the relatively flexible structure used may have constrained the views participants were able to express.

HEIs were approached by regional representatives of the UK Higher Education Funding Bodies, and asked to nominate participants for the workshops. The workshops were attended by 142 participants. These participants represented 60 out of the 172 higher education institutions, and 42 of the 68 subject based Units of Assessment, submitted in RAE2001. Around a quarter of the participants were administrative research managers. Possibly due to the recruiting methodology senior academics were over represented among the participants, relative to the more junior staff.

The workshop was structured around a number of tasks, with the views of participants collected as participant produced flip chart sheets used in feedback to the workshop group.

In the first task participants worked in pairs, to consider what characteristics mark out high quality research and what characteristics are important in research assessment

systems. All the characteristics suggested were recorded, and then prioritised by the participants using a system of multi-voting. This prioritised list provided a context for the work in the remainder of the workshop.

In Task 2 small groups of participants were allocated two of the four approaches to research assessment laid out in the Joint UK Funding Bodies 'Invitation to Contribute': Expert Review, Algorithms, Historical Ratings and Self Assessment. The groups were then asked to suggest the strengths or weaknesses of their approaches, and the questions they would want answered if that system were to be implemented. This exercise revealed how the participants thought about each of the approaches, and made them aware of the range of possibilities for research assessment.

In the remaining two tasks shuffled groups of participants were asked to design their ideal research assessment system, basing it on one of the approaches examined in Task 2. The participants were then asked to consider how their system could be implemented, what its weak points might be and how they hoped its use would change research culture in UK higher education.

Analysis

Our analysis sought to draw out recurring themes from across the workshops. For Tasks 1 and 2 this was done by grouping the participants' suggestions into related clusters. For Tasks 3 and 4 the 29 systems participant designed systems were compared to extract common design elements.

Key Findings

Peer Review

The overwhelming majority of the academics and research managers who took part in this study felt that research should be assessed using a system based on peer review by subject-based panels – of the twenty nine systems designed, twenty five were based on Expert Review. The participants also indicated that these panels should be informed by metrics and self-assessment, with some input from the users of research.

Transparency, Stability and Professionalism

There was a very strong desire for a system with clear rules, and transparent procedures, that were established at the outset and not modified during the assessment process. The appointment of panels and the selection of their criteria they used were thought to be critical areas for transparency. Participants in the study considered that the panels themselves should be professionalized and that there should be increased and earlier involvement of international members. They suggested that chair people from outside the subject area with more experience of facilitation should be used, and that these chair people might be paid.

Clarity of Submission

Almost half the groups were unhappy with the flexibility and lack of clarity over which staff should be submitted in the current assessment system, and one third of the groups felt that more staff should be submitted in future. In addition to reducing the scope for 'playing the system' it was felt that submission of more staff would improve the inclusiveness of the process. A few groups included other steps to make the process more inclusive and sensitive, both to researchers and to a lesser extent institutions.

Unit Breadth and Interdisciplinary Research

Almost half of the groups suggested that Units of Assessment should be broadened and reduced in number, with many hoping that this would help the assessment of inter-, multi- and trans-disciplinary research. Other mechanisms for improving assessment of interdisciplinary research were suggested including allowing panels to call on - or second - external expertise.

Frequency

Around half of the groups who addressed the issue of frequency recommended that the research assessment cycle should be extended, but in order to retain dynamism some suggested a light touch 'interim' assessment should be added at the halfway time point.

Agreement between Disciplines

The most important characteristics of high quality research were seen as rigour; international recognition; originality; and the idea that the best research sets the agenda for new fields of investigation. There was general agreement the importance of these characteristics by participants from different disciplines and academic roles – although absolute ranking varied.

There was also broad agreement across disciplines about the most important characteristics for a research assessment system; however, researchers from Medicine, Science and Engineering placed a greater importance on peer review, while their colleagues in the Arts, Humanities and Social Sciences felt subject related flexibility in the assessment system was more important.

Comparability and Flexibility

Participants ranked both comparability of assessments between subjects and methodological sensitivity to the subject being assessed very highly when considering characteristics of research assessment systems. Despite this, when designing research assessment systems almost half the groups suggested that panels should be given more autonomy in developing their criteria and assessment methods.

Acceptance of Burden

Although most participants were keen to avoid a system that was onerous, they appreciated that any system capable of providing the necessary fairness and would be relatively time and labour intensive. Given this realisation, it was felt that the system should provide more useful feedback to the participants to help them improve and develop their research.

Communication

Listening to participant discussions about the current research assessment system, it became clear that whatever new system is adopted, the funding councils will need to put in place programmes to engage the academics in the system's development and explain its final structure and processes.

Introduction

The Research Assessment Exercise (RAE) aims to assess the quality of research in UK universities. This information is used by the UK higher education funding bodies (HEFCE, SHEFC, HEFCW, DELNI[2]) to distribute public funds for research in a manner that rewards excellence. This mechanism aims to ensure that the infrastructure of universities and colleges carrying out the best research in the UK is protected and developed. The RAE was first conducted in 1986 and has taken place every four to five years since then, with the most recent exercise, RAE2001, having taken place in 2001.

The RAE assesses the quality of research across all disciplines – from Art History to Theoretical Physics. The research base is split into subject based Units of Assessment (UoAs), with all the submissions from institutions in each UoA being examined by a panel of subject experts. These panels rate each submission on a standard seven point scale that runs 1, 2, 3b, 3a, 4, 5, 5* in order of increasing excellence. This rating is determined by the amount of work judged to reach 'national' and 'international' levels of excellence.

The outcomes of the RAE are published and help to determine the allocation of approximately £1 billion per year of public funding to support research. Furthermore, the outcomes provide public information on the quality of research in universities and colleges throughout the UK, and are often quoted and combined into league tables by the media. As judged by the RAE, the quality of research in the UK has improved dramatically over the past decade.

RAE2001 assessed over 2,400 submissions and examined over 150,000 publications. The exercise showed a substantial increase in grading from RAE1996. In 2001 63% of research was rated as at 'national' or 'international' standard, up from 43% in 1996; and 19 universities had an average departmental score of 5 or above, up from 3 in 1996; however, despite the apparent increase in the standard of research the funding organisations were unable to provide the estimated extra £170 million of funding required to reward this improvement. Partly because of this, many voices in academia and elsewhere started to question whether the amount of effort invested in preparing for, and carrying out, the assessment process was worthwhile[3][4].

---

[2] HEFCE, The Higher Education Funding Council for England; SHEFC, The Scottish Higher Education Funding Council; HEFCW, The Higher Education Funding Council for Wales; DELNI, Department of Education and Learning for Northern Ireland

[3] MPs Signal Radical RAE Overhaul, 26 April 2002, C Davis, Times Higher Education Supplement

[4] More Join RAE Lynch Mob, 17 May 2002, S Farrar, Times Higher Education Supplement

In June 2002 the UK Higher Education Funding Bodies announced their intention to review the RAE process to ensure it was still appropriate and effective given the apparent improvement in the UK research base described above. A review steering committee lead by Sir Gareth Roberts[1] was appointed in October 2002 and tasked with investigating different approaches to the definition and evaluation of research quality, while drawing lessons from both of the 2001 RAE and other models of research assessment. Reporting in April 2003 the review will detail a number of models of research assessment, and indicate the circumstances under which particular models would be most appropriate. Sir Gareth Roberts' report will be delivered to the chairmen and chief executives of the UK's Higher Education Funding Bodies. The terms of reference of the review committee explicitly stated that the 'Dual Support System' of funding, in which research infrastructure is supported by 'Quality Related' money from the funding councils, and complemented by research grants from Research Councils and other sources, will continue and that this 'Quality Related' funding will continue to be allocated on the basis of excellence.

Three streams of evidence fed into the Review of Research Assessment: a series of open 'Town Meetings', that were held across country in late 2002 and early 2003; an online request for submissions from stakeholders and individuals and a series of nine facilitated workshops held across the UK in December 2002. This report summarises the outcomes of the workshops, which were attended by 142 academics and research managers from around the UK, and addressed the question of how research in UK Higher Education Institutions (HEIs) should be assessed. Because of the terms of reference of the review, the workshops – and this report – do not address the issue of 'Dual Support' and 'Quality Related' funding.

More information on the review process can be found on the Review of Research Assessment website at http://www.ra-review.ac.uk.

Aim and Mechanism of Workshops

The workshops aimed to stimulate broad and innovative thinking about the definition of quality research, and the mechanisms by which research quality could be assessed; but at the same time to provide a framework for this thinking, which would allow comparisons between groups and identification of common themes. We achieved this using a facilitated workshop structure that started by looking generally at the concept of research quality and research assessment, before focussing in on the evaluation and design of research assessment systems, finishing with reflection on the implementation, implications and the possible pitfalls of any new systems. The structure of the workshops was also designed to allow testing and validation of some of the ideas and frameworks emerging from the Review of Research Assessment

steering group. A detailed description of the workshop methodology is provided in Appendix A.

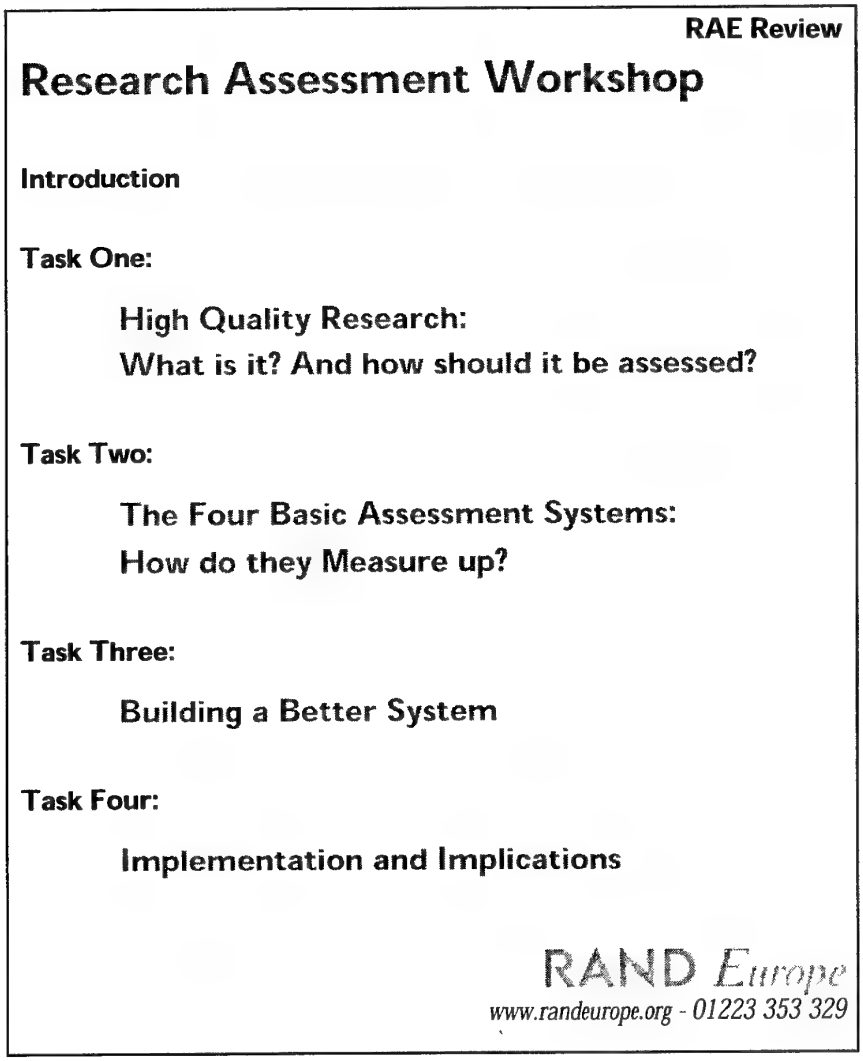


Figure 1: The workshop agenda given to participants

The workshops lasted around four hours, and were structured as a series of four tasks (see Figure 1 for the workshop agenda). Write-ups of individual workshop are included in Volume II, Annexes B, C and D, again organised by task. The tasks used a mixture of working styles: pair wise discussion, small group work and plenary reporting. The outputs were principally captured on flip charts produced by the participants, supplemented by additional notes taken during feedback sessions. The transcription of these flip charts was circulated to the participants for approval before analysis. Our role as facilitators in the workshops was to elicit ideas and gauge opinions, not to evaluate or critique comments made by participants.

The aim of this report is to present the ideas and concerns of the participants and as such they are presented without assessment by RAND Europe. Further, it would be impossible to reproduce the full richness and diversity of discussion that occurred in

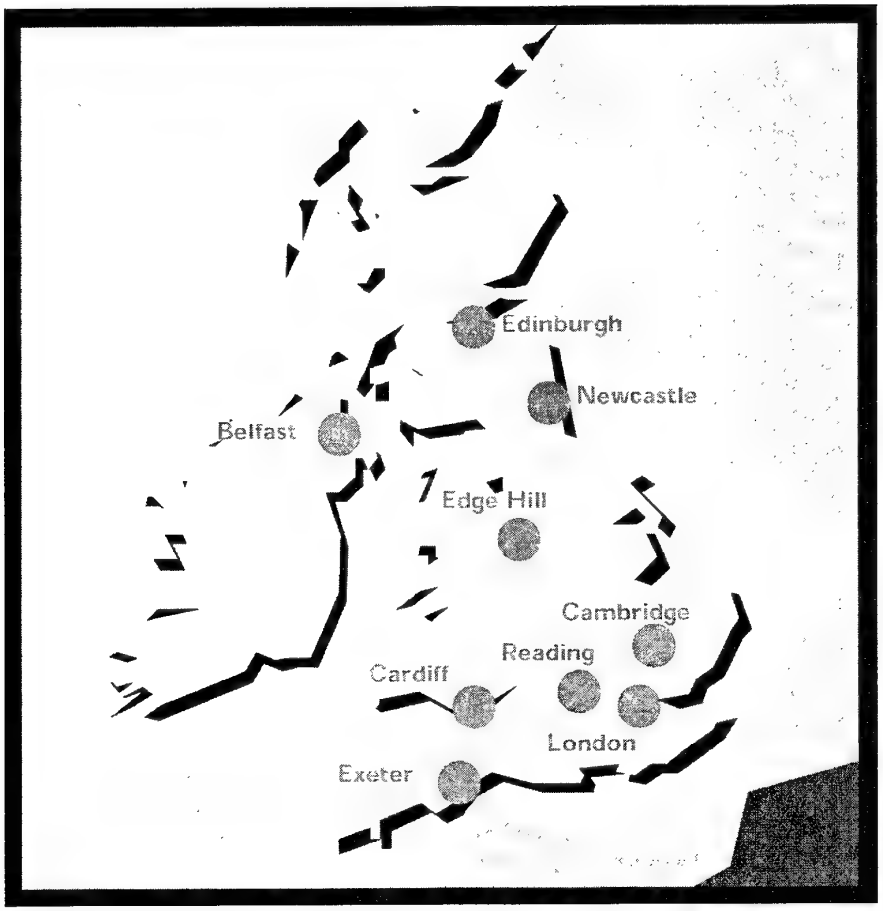


Figure 2: Map of workshop locations

the workshops in a concise report, so we have endeavoured to extract the recurring themes and opinions that emerged from the workshops. The remainder of the introduction describes the participant characteristics, and the rest of the report is structured to mirror the workshop: the next chapter reports on Task 1 – *High quality research: what is it? And how should it be assessed?* - describing what we did, how we analysed the data and the results. Chapter 2 reports on Task 2 which evaluated four basic approaches to assessment and the final chapter details Tasks 3 and 4, describing the development and implementation of alternative assessment systems.

We ran nine workshops throughout the UK. They were hosted by Cambridge University, Cardiff University, The Department of Education and Learning for Northern Ireland, Edge Hill College of Higher Education, Edinburgh University, Exeter University, University College London, Newcastle University and Reading University (see Figure 2). Participants were invited by the funding organisations' regional representatives, in consultation with local Higher Educations Institutions (HEIs). Initially, we planned to run two distinct types of workshop – those attended by participants with similar roles but from a range of subjects and those with completely mixed attendance; however, due to time constraints of invitation this did not occur.

The workshops were attended by a total of 142 participants from 60 institutions, which represented over one third of the institutions which submitted to RAE2001 (a full list of institutions is provided in Table 1, Annex A, Volume II). These participants represented a wide range of academic seniority from research fellows and lecturers to senior academics, research managers and Pro Vice Chancellors. Approximately a quarter of the workshop participants were administrative research managers. There was also a broad range of subject representation with 42 of the 68 subject based Units of Assessment represented from across the sciences, arts and humanities (see Figure 1 and Table 2, Annex A, Volume II).

From questionnaires completed by all participants we were able to analyse their characteristics, by gender, role, research field and associated 2001 RAE grade. For role we split participants into four categories:

1) research managers;
2) senior academics with large management responsibility (professor, head of department, reader);
3) lecturers or senior lecturers with teaching and research responsibilities (principal lecturer, senior lecturer, lecturer); and
4) research fellows with near exclusive research responsibilities.

For research active participants (ie, all except research managers) we also analysed their broad field of research:

1) medicine, science, engineering;
2) social science (including business and finance); and
3) arts and humanities.

Participant breakdown by role and gender is shown in Figure 3 and more details can be found in Table 3, Annex A, Volume II. When broken down by gender, research field and RAE2001 rating the participants who attended were broadly similar to the mix of academics submitted to RAE2001. Senior academics were over represented at the workshop, possibly due to the recruiting methodology, and lecturers and research fellows with little teaching responsibility were significantly under represented. This problem was particularly acute in the Social Sciences with no research fellows among the participants, and only one research fellow from the Arts and Humanities.

Of the three fields of research the largest representation was from Medicines, Sciences and Engineering, accounting for half the research active participants (Figure 4). The remainder of the participants were evenly split between Social Sciences and Arts and Humanities. More detail on the subject division of research active participants can be found in Table 4, Annex A, Volume II.

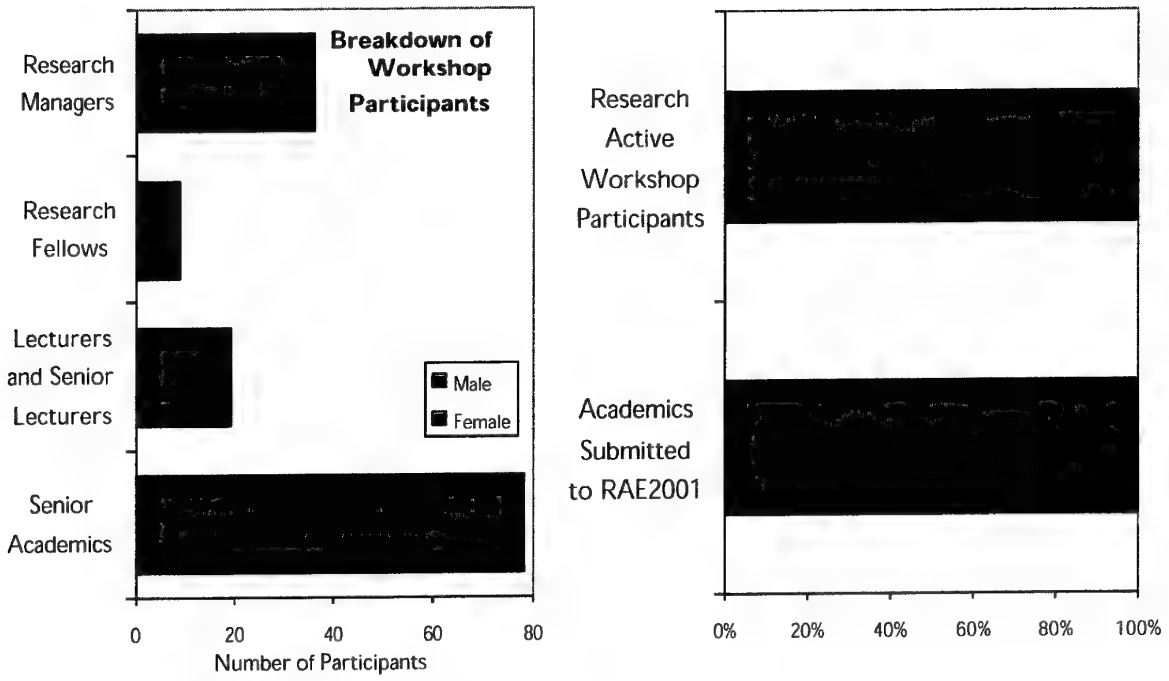


Figure 3: Workshop participants and academics submitted to RAE2001 by role and gender

Figure 5 shows the associated RAE2001 rating of the research active participants. Only three of the participants had not been entered in RAE2001, possibly reflecting the low number of junior researchers and lecturers who attended the workshops. All the RAE grades except '1' were represented among the researchers, with most of the participants coming from departments that had be awarded one of the top three grades in the RAE.

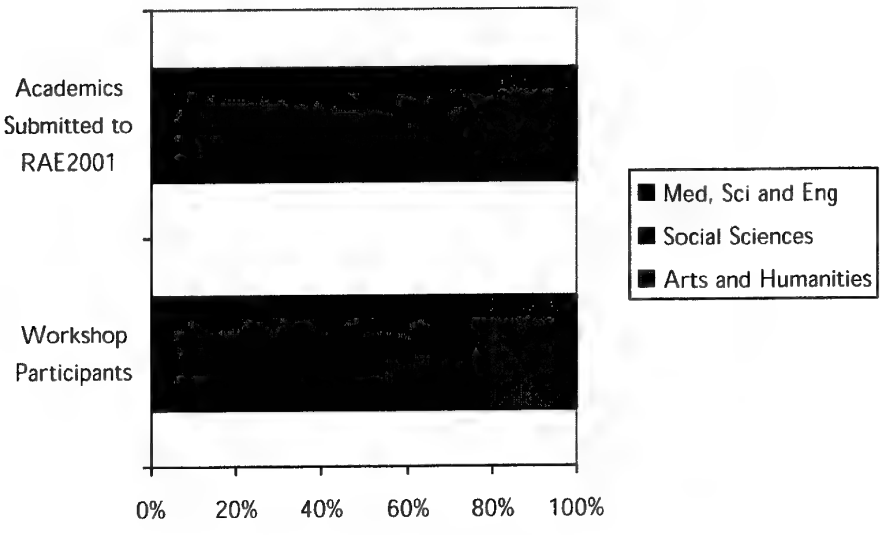


Figure 4: Comparison of fields of academics submitted to RAE2001 and research active workshop participants

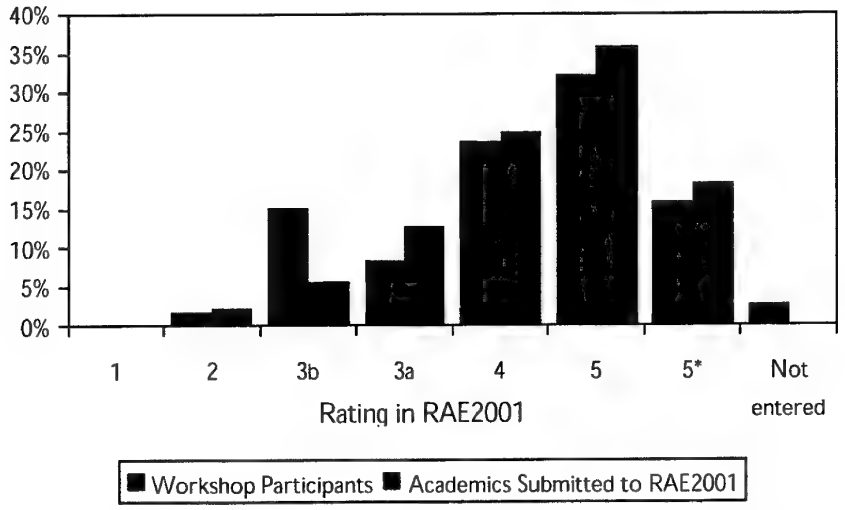


Figure 5: Comparison of RAE ratings of academics submitted to RAE2001 and research active workshop participants

Task 1 - High Quality Research: What is it? And how should it be assessed?

The Task

The first task was designed to stimulate the participants into thinking broadly about research assessment and act as an 'ice breaker'; the beginning of one workshop is shown in Figure 6. Participants were asked to work in pairs and identify five characteristics of high quality research and five characteristics they would like to see in a research assessment system.

The pairs then introduced one another, and suggested one characteristic of high quality research and one characteristic of a research assessment system. The characteristics were captured onto large hexagonal sticky labels and prioritised using a system of multi-voting.

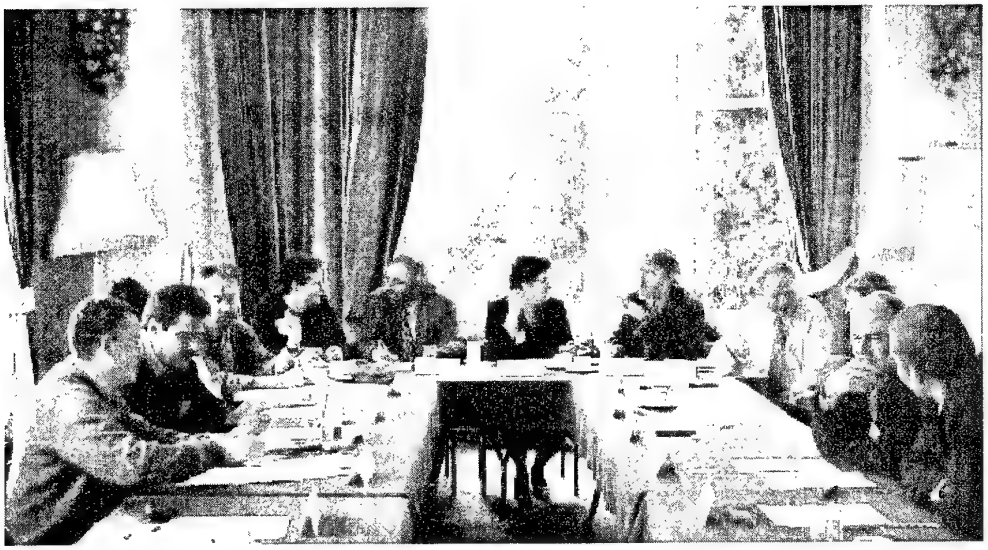


Figure 6: Discussing the characteristics of high quality and research assessment systems in Exeter.

Each participant was given five votes which could be allocated as they saw fit: if they thought one characteristic was of over riding concern they could give it all five votes; alternatively they could allocate each vote to a different characteristic; or any other combination. An example of the voting process is shown in Figure 7. A set of prioritised characteristics is shown in Figure 8. A complete list of all the characteristics suggested, and the number of votes each received can be found in Table 1, Annex B, Volume II.

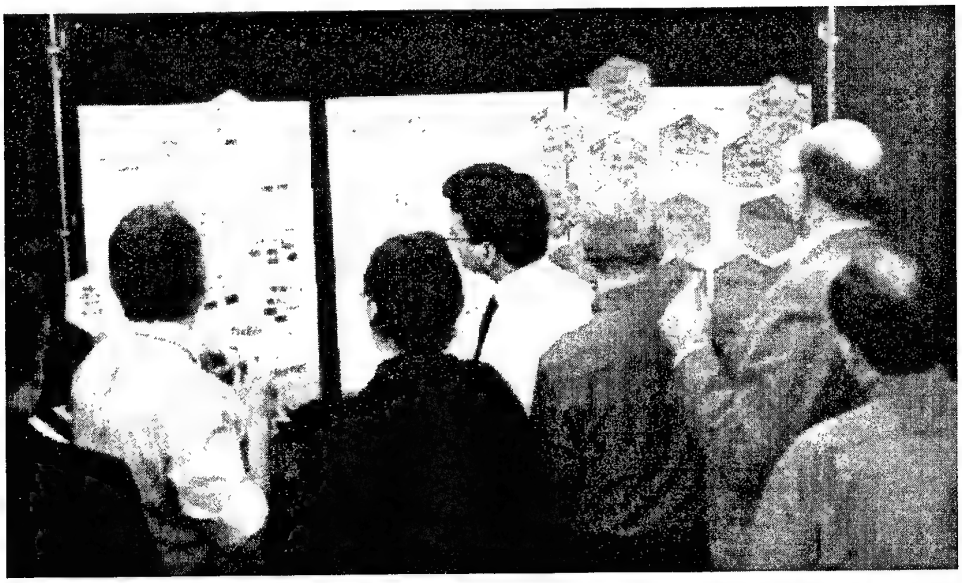


Figure 7: The voting process in Newcastle

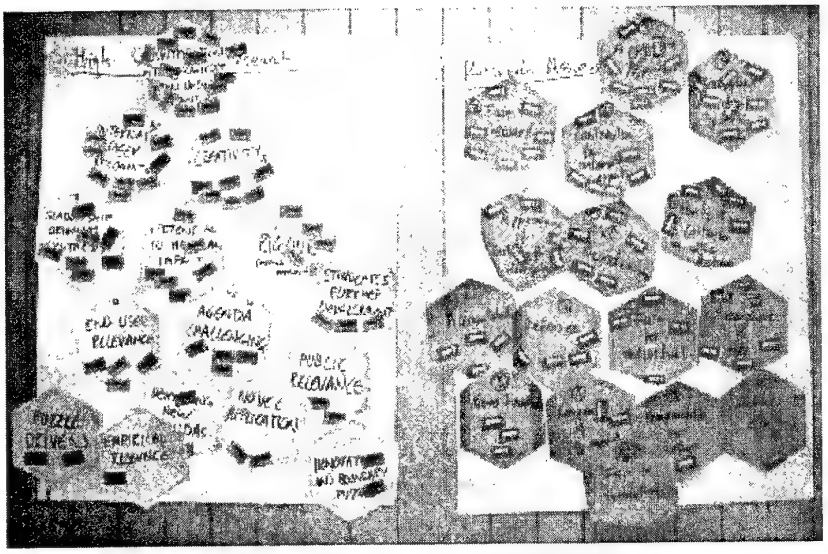


Figure 8: The voting declaration in Reading

Analysis

The multi-voting system of prioritisation was a simple but useful way of identify what the workshop groups considered the most important characteristics. In order to compare across workshops we grouped similar characteristics into clusters. Some characteristics could be very easily clustered as identical or near-identical wording was suggested by participants at more than one workshop; for example 'Peer recognition' was suggested as an important characteristic at four workshops. Other characteristics suggested at different workshops were closely related but not identical. For example 'Methodological rigour' was suggested at two workshops, but we also judged 'Rigour (methods and approach)'; 'Rigorous'; and 'Depth' to be part of the same cluster.

Given that this clustering is, by definition, subjective we have provided a complete list of characteristics and their clusters in Tables 1 and 2, Annex B, Volume II.

Results

Figures 9 and 10 show the top fifteen clusters for both high quality research and research assessment systems. The number of 'independent mentions' is also shown on the graph. This is the number of workshops at which characteristics included in the cluster were mentioned; for example, if a cluster has four independent mentions, then characteristics in that cluster were suggested at four workshops – although it may have been mentioned more than once in some of those workshops.

In addition to calculating the overall voting patterns, we also analysed voting by role of participant using the categories described in the Aims and Mechanism of Workshops section of the Introduction. Figures 11 and 12 show the top ten clusters along with the rank achieved within each subgroup of participants. As noted in the Aims and Mechanism of Workshops, some subgroups were underrepresented and therefore their results should not be over-interpreted. For example, the cluster named 'Scholarship', as a characteristic of high quality research, was only mentioned at three workshops, but scored very highly when it was suggested; however, only four (less than a quarter) of the lecturers and senior lecturers will have had the opportunity to vote for the characteristic, and no research fellows were present at any of the workshops when 'Scholarship' was mentioned.

Characteristics of High Quality Research

The concept of defining the research agenda, by framing new research questions and advancing a field into new areas was seen as the most important characteristic of high quality research. The level of rigour in the research methodology and originality of the ideas were also seen as very important, along with the concept of international recognition.

High quality publication, peer recognition, utility and academic impact, also made the top ten. The idea that research should have a long lasting impact and that value for money should be assessed did not make it into the top ten, but were present in the top fifteen characteristics.

Analysing participants by gender there was little disagreement between the rankings other than a suggestion that women scored utility higher than their male colleagues, and did not rate international recognition as highly. Looking across the different roles the top five characteristics were similar, although the order in which they fell changed. The largest difference was that research managers were more concerned about the level of peer recognition, and less concerned about the rigour with which the research was carried out. Across the disciplines there was broad agreement about the top three

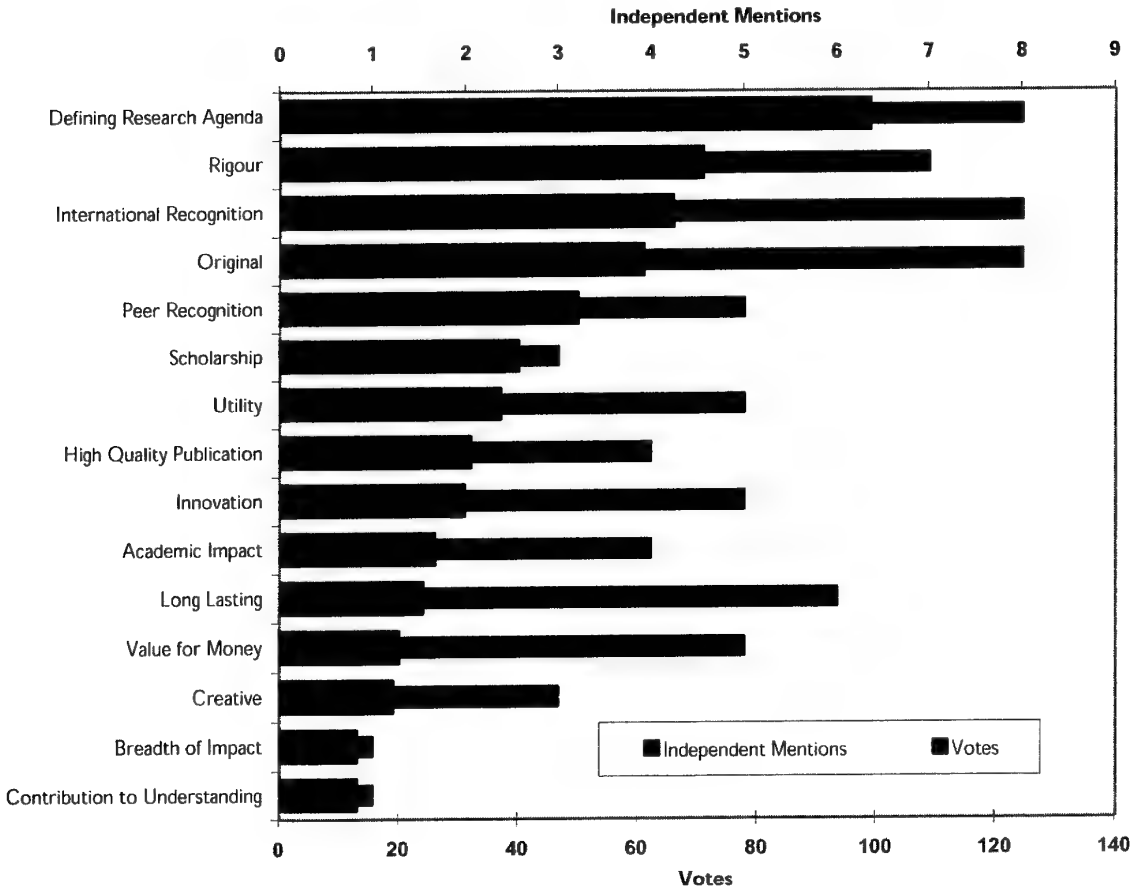
characteristics, although the arts and humanities placed less emphasis on the level of international recognition and more on originality. There was no difference in the top three ranking for departments that scored either 5 or 5* from the whole community.

Characteristics of Research Assessment Systems

The importance of transparency in both the rules and processes of a research assessment system was clearly rated as the most important characteristic, receiving almost twice as many votes as any other characteristic and coming first in every sub-group ranking other than 'lecturers and senior lecturers'. Peer recognition was also seen as an important factor in an assessment system. The other characteristics completing the top six were: appropriateness to discipline and comparability, as well as fairness and a desire that the system should not be burdensome. Although worries about game playing later emerged as a notable concern, the idea that a system should prevent game playing only just made it into the top ten of desirable characteristics.

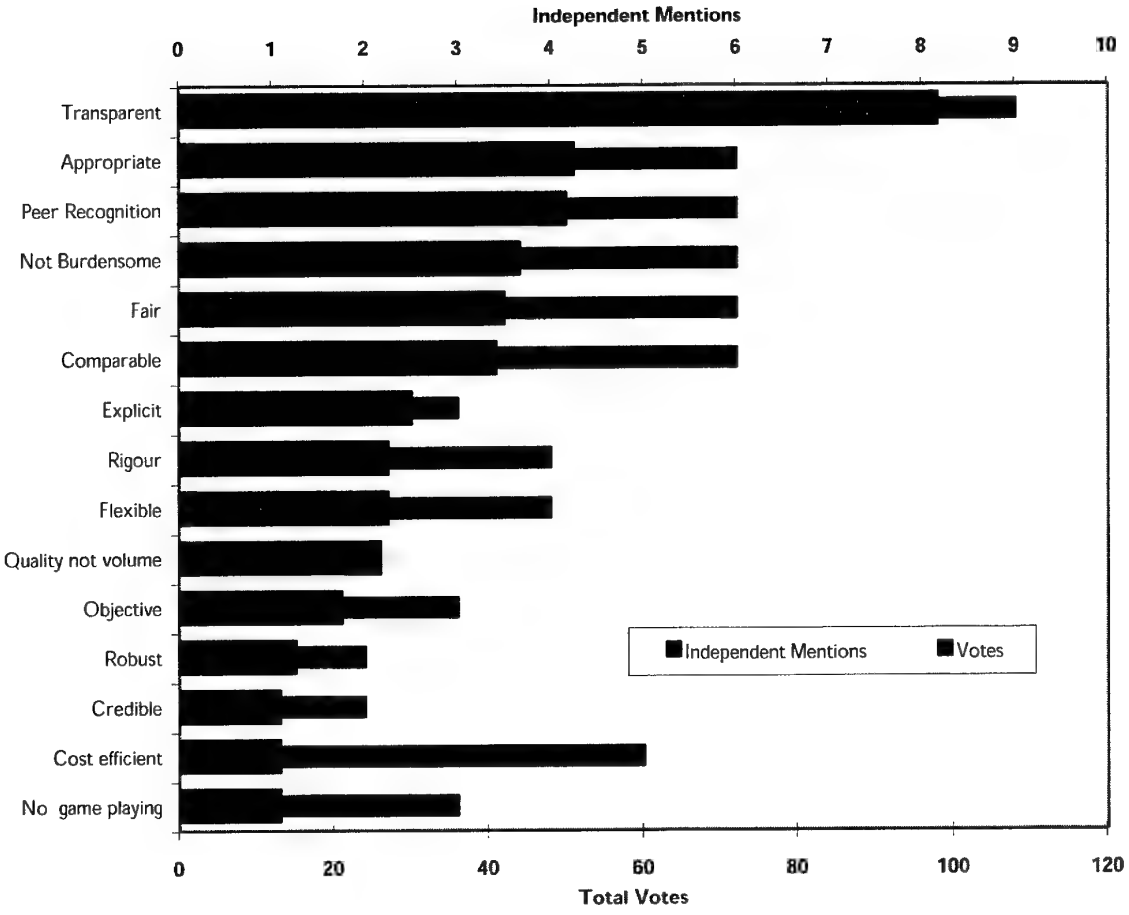
Dividing the group by gender there was general agreement on the four top ranked characteristics. There were some differences between those with different roles, with a suggestion that senior academics ranked fairness and comparability between subjects higher than the desire for low burden. Lecturers and senior lecturers disagreed placing more emphasis on a low burden system with most emphasis on peer review. Research Fellows thought it was very important to rate quality not quantity, but were not overly concerned by the burden of the system. Research managers tended to rate fairness higher than other groups, but had less concern for subject appropriateness of the system.

Looking at different disciplines there was much more emphasis placed on peer recognition in Medicine, Science and Engineering than there was in the other two disciplines. Instead, Social Sciences placed more importance in the use of explicit criteria for assessment and Arts and Humanities on the rigour of the assessment procedure; both disciplines felt that flexibility was much more important than those from Medicine, Science and Engineering.



Cluster	Explanation
Defining Research Agenda	Moving the field forward, raising new research questions and stimulating debate
Rigour	Methodological rigour and depth
International Recognition	International impact and recognition by international academic peers
Original	Originality and novelty of findings and methodology
Peer Recognition	Recognition by academic peers
Scholarship	Advancing and adding to the body of knowledge
Utility	Relevance for and impact on end users of research
High Quality Publication	Publication of research findings in high quality, high impact outlets
Innovation	Innovative and boundary pushing
Academic Impact	Academic influence and impact
Long Lasting	Longevity of impact
Value for Money	Value and productivity
Creative	Creativity and imagination
Breadth of Impact	Breadth of impact (single characteristic cluster)
Contrib to Understanding	Contribution to knowledge and understanding (single characteristic cluster)

Figure 9: Total Votes and Number of Indepencent Mentions for Top Fifteen Clustered Characteristics of High Quality Research



Cluster	Explanation
Transparent	Transparency of process and rules
Appropriate	Appropriate for and sensitive to the discipline being assessed
Peer Recognition	Impact on and recognition by academic peers
Not Burdensome	Mechanics of assessment are not burdensome or onerous
Fair	A system that is fair to both individuals and institutions
Comparable	Producing results that are comparable across subjects and disciplines
Explicit	Clear rules and criteria for assessment
Rigour	Rigorous system of assessment
Flexible	A responsive system that is flexible with regard to subject and institution
Quality not volume	Assess the quality of research and not the quantity produced
Objective	Objective assessment
Robust	Robust method of assessment
Credible	That the system is broadly acceptable and has credibility
Cost efficient	An efficient and cost effective system
No game playing	Scope for playing the system and gamemanship should be minimised

Figure 10: Total Votes and Number of Independent Mentions for Top Fifteen Clustered Characteristics of Research Assessment Systems

	Overall Ranking	Gender		Role				Subject of Research Active Participants			RAE Rating
		F	M	Senior Academics	Lecturers and Senior Lecturers	Research Fellows	Research Managers	Med, Sci and Eng	Social Sciences	Arts and Humanities	5 and 5* Depts
Participants	**139**	**38**	**101**	**75**	**19**	**9**	**36**	**20**	**25**	**58**	**50**
Cluster											
Defining the Research Agenda	**1**	**1**	**1**	**2**	**2**	**5**	**1**	**3**	**1**	**2**	**1**
Rigour	**2**	**2**	**3**	**1**	**5**	**2**	**6**	**1**	**2**	**3**	**2**
International Recognition	**3**	**5**	**2**	**3**	**2**	**5**	**3**	**1**	**3**	**5**	**3**
Original	4	3	4	5	1	2	4	7	4	1	3
Peer Recognition	5	6	5	6	9	7	2	6	8	8	9
Scholarship	6	7	6	4	7	-	9	5	6	5	5
Utility	7	4	9	10	2	7	5	8	4	8	11
High Quality Publication	8	10	6	7	11	1	11	4	14	13	7
Innovation	9	8	8	10	9	7	7	10	21	4	10
Academic Impact	10	10	9	8	13	10	8	10	10	13	10
Long Lasting	11	10	11	10	13	2	14	9	12	8	6
Value for Money	12	13	12	16	7	10	10	12	12	13	16
Creative	13	9	13	9	11	-	15	16	7	5	8

Figure 11: Rank Order Characteristics of High Quality Research for Different Groups of Participants

	Overall Ranking	Gender		Role				Subject of Research Active Participants			RAE Rating
		F	M	Senior Academics	Lecturers and Senior Lecturers	Research Fellows	Research Managers	Med, Sci and Eng	Social Sciences	Arts and Humanities	5 and 5* Departments
Participants	**139**	**38**	**101**	**75**	**19**	**9**	**36**	**20**	**25**	**58**	**50**
Cluster											
Transparent	1	1	1	1	4	1	1	1	1	1	1
Appropriate	2	4	2	2	5	5	7	3	2	2	3
Peer Recognition	3	2	3	4	1	7	2	2	7	11	2
Not Burdensome	4	3	5	6	2	14	2	6	6	7	4
Fair	5	8	4	3	12	14	4	4	5	5	4
Comparable	6	4	6	4	12	3	6	5	3	7	10
Explicit	7	12	7	7	7	7	8	8	7	11	12
Rigour	8	7	8	9	4	4	7	13	9	3	9
Flexible	8	4	10	8	9	3	25	10	4	3	6
Quality not volume	9	8	8	10	7	2	10	9	14	5	7

Figure 12: Rank Order of Characteristics of Assessment System for Different Groups

Task 2 - The Four Basic Assessment Systems: How do they Measure up?

The Task

In Task 1 participants considered the issue of research excellence and the important characteristics that a research assessment system should possess. The next task focused on looking at the four models of assessment proposed by the review body in their 'Invitation to Contribute'[5]. These four systems were:

- **Algorithms (based entirely on quantitative metrics):** A system based purely on an algorithm for combining metrics. Such a system would be automatic – leaving no room for subjective assessment. Various metrics could be included: measures of reputation based on surveys; external research income; bibliometric measures (publications or citations); research student numbers (or completions); and measures of financial sustainability.
- **Expert Review (including Peer Review):** A system in which experts make a professional judgement on the performance of individuals or groups, over a specified cycle, and/or their likely performance in the future. The groups could be research groups, departments or consortia. Assessment may be undertaken entirely by peers or may incorporate other experts such as representatives of user groups, lay people and financial experts.
- **Historical Ratings:** A system in which ratings of groups/departments/universities are determined entirely by their performance in the past. Research would, in effect, be presumed to be strongest in those departments or institutions with the strongest track record. Various measures could be used to determine past performance in addition to RAE rating, such as amount of research funds attracted.
- **Self Assessment:** In this assessment system, institutions, departments or individuals assess themselves. A proportion of the assessments are reviewed in detail. Although the assessment is made internally, external assessors could challenge the self-assessment.

For this task we split the participants into four groups. Two groups was asked to consider 'Expert Review' and 'Algorithms'; one of these groups was asked to suggest the good points of these two systems and the other group the bad points. The two

[5] http://www.ra-review.ac.uk/invite.asp

other groups were given 'Historical Ratings' and 'Self Assessment' to examine, again with one group suggesting the good points of the two systems and the other suggesting bad points. In addition to asking groups to detail the good and bad points we asked them to suggest questions they would want answered about each system where it to be adopted as the method of research assessment.

By forcing participants to look at systems from a perspective that they would not normally hold – by making them come up with good points for a system they personally did not like – we hoped to broaden the awareness of strengths and weaknesses of each system. An example of the flip charts produced at a workshop is shown in Figure 13. Any good or bad points that had been missed by the group examining a particular system could be suggested by other workshop participants during plenary feedback at the end of the task, an example of which is shown in Figure 14.

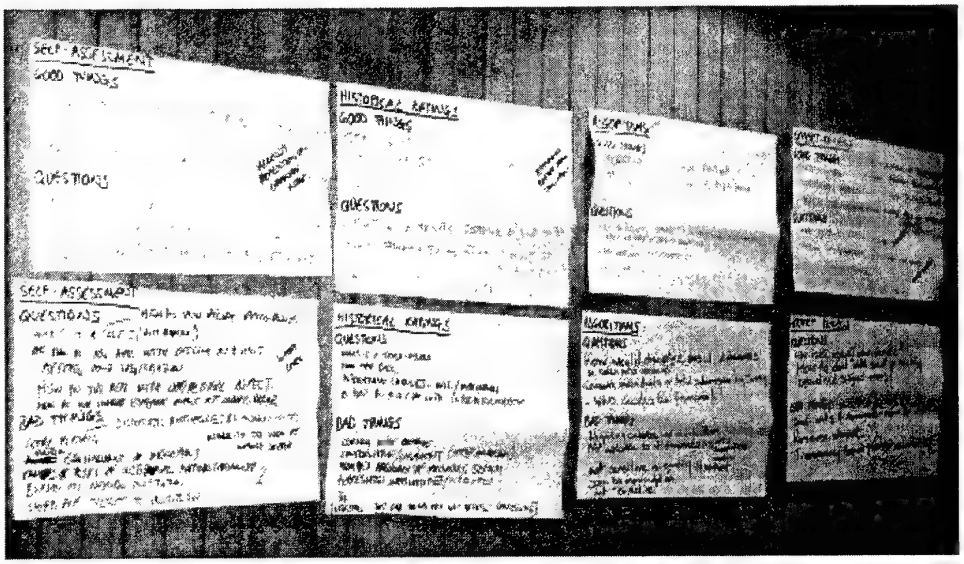


Figure 13: The good points, the bad points and questions from Reading

Analysis

In total the groups produced over 500 good points, bad points and questions. In order to highlight those ideas that are uppermost in the community's mind we have grouped the suggestions into themes[6], and we report those themes that came up in more than one workshop along with a list of the workshops in which they were mentioned. A complete lists of the good points, bad points and questions organised by workshop can be found in Annex B, Volume II.

[6] See appendix A for details.

Results

Many of the good and bad points identified by the participants were similar to those covered in the 'Notes for Facilitators' produced by the funding organisations[7]; however, the participants also suggested issues that were not covered and their responses give an insight into how the different methods are viewed by the academic community.

Algorithms (Box 1)

It was suggested that an algorithmic system had the merits of providing a transparent and easily audited process, and that it was likely to be cheaper and less burdensome; however, the groups were divided on whether such a system could be considered objective or whether this apparent objectivity was spurious. Although the possibility of different algorithms for different subject areas was considered an advantage, there was concern about whether this would lead to a lack of comparability across subjects. There was also a worry that it might be impossible to derive suitable metrics of quality for some fields, such as Arts and Humanities.

There was a widespread feeling that metrics could not truly identify quality and a worry that researchers would end up chasing the metrics rather than concentrating on the highest quality research.

The most common questions about algorithmic systems regarded how the inputs to the algorithm would be selected and combined, and about who would make these decisions. Issues of variation in algorithms across subjects and how to ensure comparability also resurfaced as questions.

Figure 14: Taking plenary feedback on Task 2

[7] http://www.ra-review.ac.uk/invite.asp

Bad Points of Algorithms

Not appropriate for all disciplines (1, 2, 5, 6, 7, 8 & 9)

The problems of finding appropriate algorithms in some subject areas, especially in humanities and arts subjects.

Spurious objectivity (3, 4, 5, 6, 8 & 9)

That algorithms were not objective, although they were sometimes considered to be so.

Quality blindness (2, 4, 5, 6, 8 & 9)

A feeling that many, or all, available metrics were bad proxies for quality.

Not comparable across subjects (1, 3, 4, 5 & 9)

The difficulty of finding algorithms and metrics that would be comparable across the whole of academia.

Open to game playing (1, 6 & 8)

Disadvantages younger researchers (6, 5)

Encourages inappropriate focus (3, 4)

Researchers focus on chasing the metrics and not on high quality research.

Reinforces status quo (2 & 5)

Crude assessment (7)

Good Points of Algorithms

Objective (1, 2, 3, 6, 8, 9)

Transparent (3, 4, 5, 7, 8)

Cheap and cost effective (3, 4, 5, 6, 7)

Numerical output is easily linked to funding (8, 9)

Flexibility with different algorithms for different subjects (6, 9)

Light touch (1, 9)

That assessment would not be burdensome

Simple (8, 4)

Questions of Algorithms

Subject comparability (1, 3, 4, 6, 8, 9)

Would the same algorithm be used across all subjects and how could comparisons between subjects be made?

Weighting of measures (1, 2, 5, 7, 8, 9)

How does the algorithm combine individual measures?

How to recognise new areas and research innovation? (1, 5, 7)

Credibility and validation of algorithms (1, 4, 7)

How is quality measured? (2, 4, 5)

Box 1: A summary of the Good and Bad Points, and Questions about Algorithmic Systems suggested by the participants (Numbers indicate the workshops in which a theme was mentioned)

Historical Ratings (Box 2)

The clearest message about Historical Ratings was their tendency to preserve the status quo. This could be seen as an advantage or disadvantage depending on the participants' opinion of the status quo. Echoing this, the biggest question about Historical Ratings was how they could deal with change, both in subject areas and institutions. Within these over arching areas there were issues of how Historical Ratings could provide motivation for improvement and/or incorporate some forward looking element. On the positive side it was felt that a system based on historical ratings would have the advantage of being very light touch and that it would focus on the tangible achievements of institutions rather than nebulous future plans.

Bad Points of Historical Ratings

Cannot cope with change (2, 3, 5, 6, 7, 8, 9)

Change either within institutions or in emerging subject areas.

Inhibits change and perpetuates inequalities (2, 3, 4, 5, 8, 9)
Encourages complacency (5, 7, 8, 9)
Does not motivate improvement (3, 5, 7)
Purely retrospective (2, 6, 9)
Lack of credibility (1, 2)

Good Points of Historical Ratings

Preserves status quo and stability aids planning (2, 3, 6, 7, 8, 9)
Inexpensive (2, 3, 4, 6, 8, 9)
Light touch (1, 2, 7)
Transparent (6, 7, 9)
Looks at track record, not ambition (5, 8)
Simplicity (1, 4)

Questions about Historical Ratings

How far back does assessment look? (1, 2, 3, 6, 8, 9)
How do you handle personnel changes and new departments? (3, 4, 6, 7, 8)
How do you encourage new subject areas? (2, 3, 4, 6, 7)
How do you measure trends? (4, 6)
Does it reflect current situation? (4, 9)
What measures will be used? (1, 3, 6, 8)
Where is incentive for improvement? (4, 5, 6, 7, 9)

Box 2: A summary of the Good and Bad Points, and Questions about Historical Ratings suggested by the participants (Numbers indicate the workshops in which a theme was mentioned)

Reflecting earlier discussion of algorithms, a common question was which historical performance indicators would be used for the assessment. There was also a desire to

know how far into the past the assessments would look and what aspects of historical performance they would examine.

Expert Review (Box 3)

When considering Expert Review it was intriguing to note, as one of the participants did, that although it was the system they were happiest with, it was the one about which they could identify least good features and most bad features. A widely held view was that Expert Review, as peer review, benefited from wide acceptance and familiarity within the academic community. It was also suggested that Expert Review was the best way to bring the necessary reviewing expertise to bear on research assessment. Many of the questions dealt with how to select the experts, what range of expertise they should have, and whether experts with the necessary expertise (especially to assess interdisciplinary research) were available.

Although there had been discussion of whether Algorithms were objective, there was general agreement that Expert Review was subjective; however, this could be seen as both a strength and a weakness. Subjectivity was seen as a major weakness laying the system open to bias and political manipulation, but it was also identified as a strength, providing an additional subtlety of assessment that was absent from more mechanistic systems.

Other disadvantages of Expert Review were considered to be its lack of transparency, and its resource intensive nature. It was also noted that it had a tendency to be conservative and could stifle innovation. Worries about comparability across fields were again noted.

Self Assessment (Box 4)

Issues of trust and credibility dominated considerations of self-assessment, with the most popular questions regarding how the system would be policed and validated. This was coupled to a concern that a self-assessment based system would lack credibility both within academia and in the wider world. There was disagreement about the level of burden that a self-assessment would involve, with some groups suggesting it was a light touch option, but others of the opinion that the checking and validation necessary to prevent exaggeration and bias would make it burdensome.

On a more positive note it was suggested that a self-assessment system could be flexible and sensitive to local conditions, but the issue of how flexible was a recurring question. This related to discussions about how comparability could be ensured and who would set the criteria for self-assessment. There were also questions about the granularity of assessment: whether it would be at an individual, departmental or institutional level.

It was felt that self-assessment would have the advantage of encouraging self-reflection and promoting engagement with the assessment process. Also, as researchers are the people who know their work best they are ideally suited to assess it.

Bad Points of Expert Review

Expertise of panels (1, 2, 4, 5, 6, 9)

> The problem of ensuring panel has necessary depth and breadth of knowledge, particular in order to assess interdisciplinary research.

Resource intensive, time consuming and expensive (1, 5, 6, 8)

Subjective (3, 6, 7)

Fails to recognise innovation and tends to be conservative (2, 5)

Potential for bias within panel (1, 2, 5, 6, 7, 8, 9)

Lack of transparency (3, 4, 5, 6, 8, 9)

Inconsistency and variation between panels (1, 3, 5, 7, 9)

Good Points of Expert Review

Acceptance of peer review (1, 2, 4, 5, 6, 9)

> Peer review has the confidence of being widely known and accepted in academic community.

Can include international input (3, 8, 9)

Employs specialist knowledge of reviewers (2, 5, 6, 8, 9)

> Can pick up subtleties and non-quantitative aspects of quality.

Flexible and sensitive to discipline (7, 9)

Questions about Expert Review

How can the process be seen to be transparent? (2, 3, 4, 5, 6, 7, 8)

Can workload be managed and afforded? (3, 5, 7,)

Panel composition and selection (1, 2, 3, 4, 5, 6, 7, 8, 9)

> Who chooses the panel and who are they selected from?

How do you avoid bias? (2, 4, 7, 9)

How do you deal with very specialised areas? (5, 8)

How do you deal with interdisciplinary research? (2, 3, 5, 7)

What criteria and what data? (4, 9)

How do you ensure comparability across panels? (1, 3, 4, 6)

Box 3: A summary of the Good and Bad Points, and Questions about Expert Review suggested by the participants (Numbers indicate the workshops in which a theme was mentioned)

Bad Points of Self Assessment

Lacks credibility (1, 2, 3, 6, 9)

Could be burdensome (2, 3, 8, 9)

Game playing (2, 6, 8)

Problems of comparability (3, 5, 7)

Too inward looking (2, 6, 9)

Institutions over rate themselves (3, 5, 9)

Subjective and prone to bias (2, 9)

Has parallels with Teaching Quality Assessment system (6, 9)

Good Points of Self Assessment

Simplicity and low cost (1, 3, 4, 8)

Flexible and sensitive to local conditions (3, 5, 6, 7)

Formative and encourages reflection and responsibility (2, 5, 8, 9)

Increases engagement of researchers with process (2, 4, 5)

Questions about Self Assessment

Validation and policing (3, 4, 5, 6, 7, 8)

What granularity of assessment? (1, 5, 6, 7, 8)

> Is self-assessment by researchers, research groups, departments or institutions?

Who sets the criteria for the assessment? (2, 3, 6)

What is the 'self'? (6,8)

How is comparability ensured? (7, 9)

What evidence would be required? (2, 4)

> How would the self-assessment be validated and how would inaccurate assessments be dealt with.

Box 4: A summary of the Good and Bad Points, and Questions about Self Assessment suggested by the participants (Numbers indicate the workshops in which a theme was mentioned)

Tasks 3&4 - Building a Better System, Implementation and Implications

The Tasks

Having considered research quality and assessment, and then assessed the strengths and weaknesses of four possible approaches, we next asked small groups (of between three and seven participants) to devise their ideal research assessment system. One such group is shown in Figure 15. In Task 3 we asked the groups to design their ideal system. In Task 4 we asked them to consider how their system could be implemented; what could go wrong; and, finally to consider what changes they would like their system to produce in the research base.

To devise their research assessment system each group was asked to start with a kernel system of their own choosing – either one of the four approaches they had already considered or an entirely different one that they preferred. They then had to build on this system by adding features to enhance the good aspects and mitigate against the bad aspects identified in Task 2. The aim of this process was to devise a system that had as many of the desirable characteristics identified in Task 1 as possible.

Figure 15: Designing the ideal system in Edinburgh (Task 1 and Task 2 output in background)

In the early workshops there were a number of interesting ideas that emerged so, in order to test these concepts, we produced a prompt card showing the ideas and gave it to the groups halfway through Task 3. We told them that the ideas listed were

provided simply to stimulate thinking and they could use or discard them as they pleased. The prompt card is shown in Figure 16.

Some Previous Ideas

Sampling– using either selective or random assessment of research submissions, possibly coupled to a light touch assessment for all submissions.

Triggering – triggering an assessment through pre-defined event such as self-declaration or change in metrically assessed output.

Lay panel members – using lay members on panels, or possibly experienced lay members such as judges as chairs for panels.

Review interval – either increasing the review interval, or having a continual rolling process of review.

Funding high flyers – direct allocation of funds to high fliers in low rated departments.

Transfer Fees – providing recognition to departments that nurtured researchers who then become high flyers.

Figure 16: A prompt card of systems ideas

After the plenary feedback for Task 3, in which each group presented their system to the whole workshop group (an example flip chart is shown in Figure 17), we asked one member of each group to swap. This 'Devil's Advocate' provided each group with an external viewpoint and a critical voice, as they considered the implementation and implications of their system.

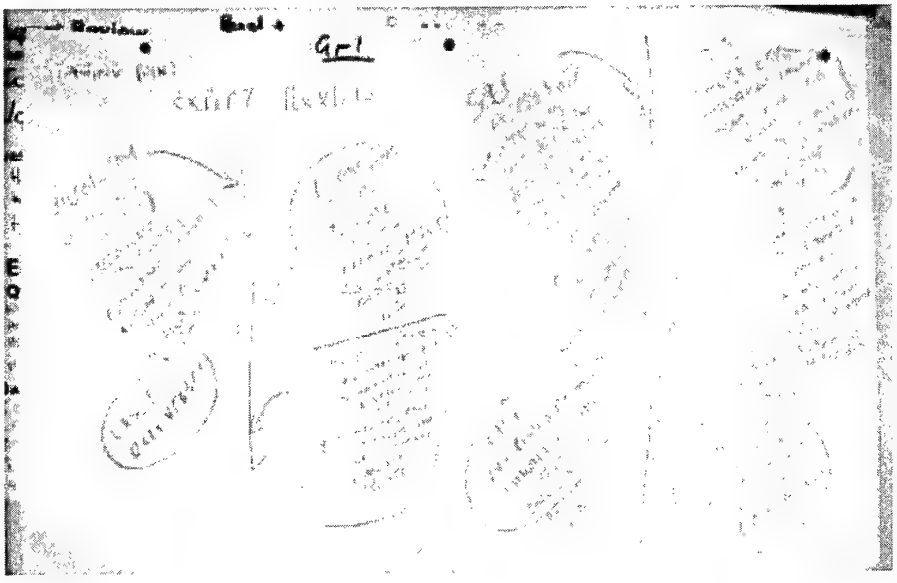


Figure 17: One of the twenty-nine research assessment systems

Analysis

There was no clear split between what groups chose to present as part of their system design (from Task 3) and the detail and fine tuning they chose to present as part of

the implementation (first part of Task 4). Consequently we have chosen to merge the reporting of the two tasks. Similarly some of the aspects that were seen as potential failures of the system (second part of Task 4) related directly to aspects identified in system design and so are reported alongside them. Aspects from the first two parts of Task 4 that did not fall neatly into aspects of system design are reported in their own section. The outcome of the final section of Task 4 is presented as the final section of the report.

Results

Preferred Kernel System

Over the course of the nine workshops, 29 groups devised their own assessment systems and 25 of these groups chose to use Expert Review as their kernel. Of the remaining four groups: two opted for systems based on Self Assessment, with Expert Review to provide rigour and consistency; one opted for a system based on Algorithms, heavily moderated by Expert Review; and one group failed to reach a decision about the base system. As most groups started with Expert Review many came up with systems that bore significant similarities to the system used for RAE2001[8]; however, all the groups modified and built on Expert Review in different ways to produce what they considered a better system. Many of the modifications they suggested fall into themes[9] and these are discussed in the next sections. Some ideas only occurred once, but are noted as we consider them to be novel enough to warrant discussion. Details of all the systems are in Annex D, Volume II. The four systems that were not based on Expert Review are included in the Annex D, Volume II as systems 1, 12, 13 and 20.

Suggested Modifications to Expert Review

Data for Panels

(Mentioned in systems 2, 3, 4, 5, 6, 7, 8, 9, 10, 14, 16, 17, 18, 21, 22, 23, 24, 25, 26, 28 & 29)

Most groups addressed the issue of what data should be provided to panels and how they should deal with it. All the groups that addressed the matter indicated that panels should identify their data requirements and most felt that panels should be allowed more autonomy in this area. This suggestion of increased panel autonomy is also covered under issues of panel breadth. In one instance it was suggested that the range of data examined by the panel should be open to consultation with the subject community. Other groups noted the rules and criteria of the process should also be open to consultation by the panel in the subject community.

---

[8] http://www.hero.ac.uk/rae

[9] See Appendix A for explanation on how themes were identified.

In addition to the data which panels currently use it was noted that research application success rates and the opinions of research users could provide valuable input into the process; however, there was disagreement about how much importance should be given to research user views, mirroring discussion of whether they should be included as panel members. Three groups suggested that it would be useful to broaden the scope of what could be counted as a research output, to include for example exhibitions. One of these groups noted that this broadening would be particularly useful in the Arts and Humanities. Another group suggested that panels could use institutional visits. Going against the general tide of collecting more data for the panels, two groups suggested that much more emphasis should be placed on pre-existing and pre-collated data from research councils and within universities to reduce the burden of the system. It was also suggested that, in order to try and avoid data overload, HEFCE could consult with the panels to produce a pick list of data sources. The panels could then choose the data sources they wished to examine. Another suggestion, raised by five groups, was to develop an efficient and easy-to-use system of electronic submission. Reflecting the level of concern about current practice, three groups specifically identified software failure as a vulnerability of their systems.

Four groups suggested that the practice of only submitting four pieces of work per academic should be dropped in favour of a system where all work was submitted and the best pieces of work highlighted. Four other groups thought that it was important that all submissions were read, to guard against assumptions based on reputation and to ensure fairness, conversely one group thought that sampling was acceptable provided at least one piece of work from each academic was read. Two groups thought that algorithmic assessment offered the possibility of reducing the workload of the panel.

Given the comments in Task 2 (Box 4) it was unsurprising there was great ambivalence towards what role self-assessment should play in providing data to panels. Of the nine groups that dealt with self-assessment, two groups thought that it should either be integrated more effectively than in the current system or scrapped altogether: 'use it or loose it' as one group put it. Other groups felt that self-assessment provided a valuable way to set the rest of the assessment - such as the published outputs and metrics - in context, and provided an opportunity to incorporate the career sensitivity that the system needed. Two groups suggested that a 'panel reviewed self-assessment' provided an opportunity for more prospective assessment than other techniques. If self-assessment was to play these roles it was seen as important that it be structured, with that structure probably set by the panel.

Two groups saw 'panel reviewed self-assessment' as playing a particularly important role in their systems. System 26 (Annex D, Volume II) proposed a panel with a more

developmental role with 'panel critiqued self-assessment' facilitating improvement. In system 18 (Annex D, Volume II) 'panel reviewed self-assessment' played a large role in the interim assessments when UoAs would review themselves against criteria set at the previous full-scale assessment.

Period of Assessment

(Mentioned in systems 2, 3, 4, 5, 9, 10, 14, 15, 16, 17, 18, 21, 22, 24, 25 & 26)

Just over half the groups considered the timing of assessments and seven of these groups thought that the current system of assessments every 5-6 years should be continued, although one thought they needed to be supplemented by interim assessments to help new UoAs.

Five groups considered that to reduce the burden of assessment the period between assessments should be increased, with the most common suggestion being an interval of 8-10 years. One problem of such a long interval between assessments was it could prevent recognition of improving UoAs, and this would make it harder for new fields to emerge. To address this problem three of the group suggested 'interim assessments': lighter touch assessments, based more heavily on metrics and/or self-assessment than the full assessments. It was felt that interim assessments required two mechanisms to be effective – a system for new improving UoAs to opt-into the system and a system to pick up declining UoAs; although self-assessment could be sufficient for the former, it was thought that some type of metric-based trigger would be needed to detect decline. In some systems this assessment trigger was used to call in a full reassessment of the UoA at the interim time point and in others it was used directly to regrade the UoA. One group suggested that the interim assessment should be concerned with assessing how the UoA has done in meeting targets set in a dialogue with the panel at the previous full assessment.

Two groups suggested that full assessments would not need to be fixed to a rigid timescale with all UoAs being assessed in the same year. One of these groups suggested that full assessments could be triggered by a change in metrics and the other suggested that there should be an algorithmic assessment every two years with UoAs able to call in a full assessment when they felt it was appropriate.

Subject Breadth of Panels

(Mentioned in systems 2, 3, 5, 6, 8, 10, 11, 14, 15, 16, 17, 19, 21, 22, 24, 26 & 28)

Thirteen groups suggested that panels should be reconstituted as broader 'supra panels', and that these should be allowed more autonomy. In providing this level of panel autonomy the groups downgraded the importance of comparability across disciplines, moving to a position that placed more emphasis on flexibility and subject appropriateness of the assessment methodology. A number of these groups then

engaged in the problem of subject specific expertise by including more sub-panels reporting to the 'supra panels'. Addressing the question of how this restructuring could be carried out, four groups suggested that it should be carried out by a top-down consultation process, although one suggested the groupings could be generated in a bottom-up manner. The volume of research work in an area was put forward as a possible method of determining panel size and breadth of subject coverage, the aim being to have a similar volume of work examined by each panel.

It was suggested that broader panels would find it easier to assess cross and interdisciplinary research as more of it would fall within the panel's purview; however, there would still be boundaries between panels. A number of mechanisms of dealing with this were suggested: five groups suggested the use of external referees, who could be nominated by the researcher; alternatively experts could be seconded onto the panel. It was further proposed that the chair of the panel be given specific responsibility for ensuring that cross and interdisciplinary research was properly assessed. At the other extreme it was suggested that responsibility be left with the submitting academics and that they be allowed to fractionate their work by submitting parts of it to different panels and be credited for it in their 'home' panel.

Panel Members and Role of Panel

(Mentioned in systems 2, 3, 5, 6, 7, 8, 9, 10, 15, 16, 17, 18, 22, 24, 25, 26, 28 & 29)

There was much discussion about panel members and many suggestions about panel composition. Six groups specifically noted that that the panel members should be selected by a transparent process that involved wide consultation with the subject community. Some thought that universities should also be able to put forward nominations for panel members and one group suggested election of panel members. Panel composition was also identified as a possible failure point by nine groups, with particular concern about on the problems of finding willing and suitable panel members. It was felt that if panels were to cover broader subject areas then panel members would need to be selected with specific attention to their breadth of knowledge. It was also suggested that a wider range of researchers should be selected for panels including younger researchers. Three groups noted that there should be a high level of 'churn' in panel membership and the role of chair should not be inherited from the previous round. One group advocated complete 'churn' for every round of assessment.

There was some disagreement between groups about whether panels should include the users of research: with three group suggesting they should; one group was firmly against including research users; and one suggested a preference for peers over non-peers. Eight groups felt that panels needed to be less subject introverted and

suggested that panels should include members from other subject panels or lay members.

Seven groups noted that the chairing of panels needed special attention and suggested that chairs could be appointed from different disciplines as this could provide more comparability between panels and prevent issues of subject self interest. One group suggested that lay chairs could be appointed for their experience of chairing rather than their subject expertise and a couple of groups considered that chairs should have a purely facilitating role. One group suggested that chairs could be elected by their panels. It was also felt that the panels' membership needed to be professionalized, possibly by payment of all members or just remuneration of international members.

International panel members received considerable attention. The eleven groups that addressed the issue suggested that international panel members needed to be fully engaged in the review process and should be involved earlier, which would allow them to provide more valuable input. It was noted that the concept of 'international' needed to be broadened beyond North America, and that the concept of 'international' has particular problems for some subject areas such as history, as one participant commented: 'French French historians have a very different view of French history to English French historians'. One group felt that the process of appointing international experts needed to be as transparent as the process for appointing UK panel members, and that if this could not be achieved then input from international experts should be scrapped altogether.

Five groups mentioned that the role of the panel should extend beyond rating the quality of research outputs and include reviewing development plans. One of these groups felt that the panel should have a permanent developmental role in the subject area, entering into a dialogue with UoAs about strategy and development.

The Rating Scale

(Mentioned in systems 2, 3, 4, 5, 7, 9, 11, 14, 15, 18, 22, 23, 24, 26 & 28)

There were many suggestions for modification of the rating scale. These suggestions seemed to arise from the concerns regarding the step changes in funding in the current system. Two groups favoured an expansion in the number of ratings beyond the current seven; however, most groups that addressed this issue preferred a move to a continuous scale in which a submission's score was the sum of the scores of the individuals included in the submission. One group suggested there should be limits on the number of high rated submissions within each UoA, to avoid a cartel situation where every submission in the UoA is awarded a high rating.

One group suggested that rather than a single rating, subsuming all the individual criteria, each submission should be evaluated against a number of different criteria to

give a profile of the submission's strengths. This was intended to allow different institutions to focus on their strengths. In order for this system to work it was suggested that the highest level of funding would have to be awarded to institutions that achieved top rating for a fraction of the criteria, for example three out of five, otherwise institutions would continue to chase top rating in all criteria. One possible problem of using excellence profiles was that if all institutions rated highly in some area they would all achieve the same overall rating. This would result in a uniformity of result and the advantages of funding concentration could be lost.

There was also discussion at two of the workshops about the appropriateness and meaning of the terms 'internationally competitive' and 'nationally competitive'. It was felt by many that research excellence was an anational concept, so the idea of national competitiveness was unhelpful. It was also thought that confusion between 'international excellence' and 'international relevance' could disadvantaged high quality research which had relevance only within the UK (eg, research into the National Health Service) or relevance only within a region (eg, tourism). Confusion over the concept of 'international excellence' was also cited as leading researchers to publish in non-UK journals, because these may be inappropriately seen as more 'international' than UK journals. These issues were also touched upon when groups considered how review panels should include an international perspective.

Fraction of Staff Returned

(Mentioned in system 4, 5, 7, 8, 9, 10, 11, 15, 17, 23, 24, 25 & 28)

All groups who dealt with this issue were unhappy with the present system, they felt that as a bare minimum there needed to be more clarity over which staff should be included in submissions, with particular attention paid to staff who are contracted partly by other organisations such as the NHS. Of the ten groups suggesting an increase in the number of staff that should be returned, one simply suggested more should be returned, five suggested that all staff should be returned and four groups pursued a middle way suggesting that at least 70% of staff should be returned or the grade of the submission should be capped. One group trying to reconcile 'the pitfalls of not including all researchers against the alternative pitfalls of submitting everyone' came up with a suggestions that staff should be returned as 'research time equivalents', ie all staff should be returned but the fraction of their time devoted to research should be detailed, an approach that was also favoured by another group.

Clarity and Transparency of Process

(Mentioned in systems 2, 3, 5, 9, 10, 11, 16, 21, 24, 25, 26, 27, 28 & 29)

There was a general feeling that the process of assessment from panel appointment to submission grading needed to be as clear and transparent as possible. There were many calls for the rules and criteria of the process to be set early – the suggestion of

five years in advance was made on several occasions – and that the rules and criteria should not be changed in the run up to the assessment (an eventuality that was identified as a possible system failure on three occasions). There was also a feeling that there should be consultation with the community to set these rules and criteria, particularly panel specific processes and criteria. Eight groups also suggested that the funding implications of the process needed to be clear from the start, this aspect is dealt with in more detail in the next section on funding.

Over one third of groups felt that the assessment process ought to provide more detailed and useful feedback, either at the end of the process with the panel highlighting areas that needed improvement to win and improved rating, or with the panel and the institution entering into a dialogue during the process so the institution is able to clarify issues noted and address concerns raised by the panel. It was observed that one reason for a lack of feedback from RAE2001 might have been the threat of legal action from the institutions against panels or even panel members. This lead to a suggestion that the funding councils needed to construct a system that was resistant to such challenges, possibly by making entry to the system by way of a contract where the funding councils agree not to change the rules of the system in return for which institutions agree not to challenge the results.

Funding

Mentioned in systems (5, 7, 9, 10, 11, 14, 16, 17, 22, 24, 26 & 27)

The way in which success in the RAE translated into funding outcomes was considered by a number of groups. There was often a view that step changes in funding, such as the ones currently caused by moving up or down a grade, needed to be avoided. To achieve this, a rating system with more subtle graduations was necessary, consequently this theme is intimately connected with proposals to modify the rating scale discussed previously.

Eight groups thought there should be clarity of outcome: institutions needed to know what the implications of success and failure were. Balancing this the groups also recognised that the outcome was dependent on the level of funding attained by the funding councils, so it was suggested that a number of funding scenarios could be presented. The possibility that funding would fail to increase in response to a demonstrated increase in research excellence was identified by five groups as a significant system failure.

Five groups considered that there should be effective mechanisms of targeting funding to successful individuals in low rated UoAs, although one of these groups considered that this might be inappropriate in subjects where the infrastructure costs were high and hence it would be inefficient to provide infrastructure for a lone (successful) individual.

One group also considered that improvement in excellence, as well as absolute excellence should be rewarded and another considered that 'Quality Related' funding should start further down the rating scale.

Use of Algorithms and Metrics

(Mentioned in systems 3, 7, 8, 9, 11, 15, 18, 25, 26, 27 & 28)

Many groups thought that metrics should be used by the panel as part of the assessment process. Two groups cautioned that these metrics should only inform and not determine the decisions of the panel. A further two groups also thought that appropriate metrics offered a way of reducing the burden of the assessment process, although bad metrics were also identified as a failure mode of the system. Half of the groups considering the use of metrics stressed that they would need to be subject specific and this was particularly true when using metrics to assess external research income. It was noted that this might be particularly true for work in the Arts; as such work was often supported in kind or with money that did not appear in the University accounts.

Considering bibliometric measures it was noted that this would require some kind of equivalence scale for alternative research outputs such a journal articles and books, and that citation impact would need to be calibrated for confounding factors such as academic seniority. One group was keen to use metrics to assess the long-term impact of researcher and another thought that research income should not be included in any metric.

Evaluation Criteria

(Mentioned in systems 4, 5, 6, 17, 18, 19 & 29)

A number of groups clarified what should be considered as research excellence by suggesting criteria that should be used for evaluating submissions. The most common criterion, flagged up by four groups, was research strategy. Two groups noted that the quality of outputs should be assessed and another two suggested that an indicator for value for money, or return on investment, needed to be developed. Evaluating research culture, and how institutions nurture and develop research staff was mentioned by two groups, and one of these wanted to develop metrics to allow easy assessment of this area. Looking at esteem indicators another group suggested that panels should produce lists of acceptable esteem indicators such as membership of royal society, journal editorships etc.

Inclusiveness

(Mentioned in systems 4, 5, 7, 8, 9 & 10)

Mechanisms to increase the inclusiveness of the system fell into two categories, those aimed at allowing wider participation of individuals in the system and those aimed at

allowing wider participation of institutions. It was felt that the current system could have negative repercussions on young researchers and those who took career breaks, and that much of this related to the perceived need to submit four research outputs. Considering institutions it was also felt that there should be no bar on smaller, or less research focused, institutions entering the system. One way to achieve this would be to provide more effective mechanisms for dealing with joint submissions, and another would be to score individuals and grade submissions by combining these scores.

Appeals System

(Mentioned in systems 9, 25, 26 & 27)

Although the idea of an appeals system or oversight panel came up explicitly in only three of the systems, it was widely supported by participants and probably derives from same unease that stimulated the suggestions for increased openness and transparency of process. These suggestions may also relate to suggestions that there should be more dialog with the panel during the assessment process. None of the groups detailed their desired appeals procedure, although, one suggested an ombudsperson role and another suggested that metrics might be used to support appeals. The downside of an appeals system was recognised by two of these groups, who worried that the system could be overwhelmed by challenges.

Transfer Fees

(Mentioned in systems 6, 9, 15 & 24)

Often young researchers spend their early career in one department or institution and then go on to achieve academic success once they have moved to another institution. To reward institutions that supported the UK research base by nurturing and developing new talent, it was suggested that better methods of rewarding the original nurturing institutions need to be incorporated into the system.

Regional Issues

(Mentioned in system 9, 10, 28 & 29)

Although it was agreed that there should be the same standards of excellence throughout the UK, it was felt that there needed to be more sensitivity in the assessment process to different modes of working. An example of this was that, in areas with smaller institutions there is more collaborative work between them, and this work should not be disadvantaged in the assessment process because of the mode of working that produced it.

Opting Out

(Mentioned in system 28 & 25)

There seemed to be a general assumption that all institutions should be assessed using the same system. Only one group suggested there might be two systems of assessment

with submissions that scored four or below, being able to opt out of rounds of the RAE if they didn't feel they were likely to improve their rating. One other group identified the emergence of a two-tier system as an undesirable outcome of any research assessment system.

Non Subject Based Units of Assessment

(Mentioned in system 19)

It was suggested by one group that the conventional approach of subject based UoAs be radically restructured and replaced with a system in which research was grouped according to its type of outputs. Four different over arching Areas of Assessment (AoAs) were identified: Formal Sciences, that produce theorems; Explanatory Sciences, that produce laws; Design Sciences, that produce technological rules and Human Sciences that produce artefacts and knowledge.

Points from Non-Expert Review Starting Points

(Systems 1, 12, 13, 20)

Four groups did not build systems based on Expert Review, although all incorporated some level of peer review. Two of the groups combined Self Assessment and Expert Review, one Algorithms and Expert Review and the fourth group failed to agree on a system of assessment. Many of the aspects incorporated into these four systems have been covered above but a brief summary of the systems and some of the remaining points are reported here.

The first Self Assessment based system (System 1, Annex D, Volume II) was structured around UoAs producing self-assessments, around 20% of which where reviewed by an expert panel. The UoAs to be reviewed would be selected by three mechanisms: by random selection; where there was a large discrepancy between the self-assessment and either current metrics; or the UoA's historical rating.

The second Self Assessment based system (System 13, Annex D, Volume II) was based on a split of funding into core and competitive funding, where the competitive funding would be allocated by review of self-assessments. One role for panel was to develop tools and frameworks for self-assessment for use by institutions. It was also suggested that excellence should be split into a number of areas, such as research outputs, research culture, infrastructure and each submission should be rated against each criteria, these ratings which would then be funded from different funding streams.

The Algorithm based system (System 20, Annex D, Volume II) was an attempt to reduce the burden of the system by increasing automation and using more quantitative measures, although it was noted that this could lead to reinforcement of the divide between subjects in which metrics can be used and those in which suitable

metrics do not exist. A model was envisaged in which 'In science judgement moderates data and in arts data moderates judgement'. It was suggested that a strength of the system was its reliance on measures based outputs.

Implementation of the New Systems

When the groups considered how their system should be implemented they often took the opportunity to provide more detail on how their system would function, or to detail how they would achieve what they had prescribed in their system design – such as transparent panel appointment procedures. All of these points have been collected into the discussion of modifications to the expert review system, what remains are a couple of general points relating to implementation process.

There was a broad consensus that implementation would depend on a top down process driven by the funding organisations. If the system called for it, this would start with wide consultation on UoA structure and then proceed with a transparent panel appointment procedure and discussion of how each panel would carry out its assessments. It was generally felt that the panels should be appointed early in the process and that they should be involved in shaping the assessment process.

What Could Go Wrong with the New Systems

The possible points of failure that the groups identified fell into three categories: those that affected the implementation of the assessment system; those relating to its operation; and those relating to the results of the exercise. Some of these points related to specific aspects of the assessment system and have been dealt with in the previous section; the remainder that relate to broader issues are presented below.

Although the groups were generally in favour of extensive consultation to agree the frameworks of assessment three groups worried that consultation overload could occur. There was also a worry that deadlines for agreement of panel membership and assessment criteria would be missed.

The issue of greatest operational concern, identified by about a third of the groups, was the new system would increase the burden on researchers distracting them from their research work. A similar number of groups were concerned about the impact of gamesmanship on their system; one aspect of this - for systems with an interim assessment - was abuse of the selection used to admit institutions to the interim assessment. Four groups identified negative impact on unsubmitted staff, and one suggested the emergence of a two-tier system (in which teaching only universities emerged) would be a failure. One group worried about inappropriate pressure being bought to bear on panel members and another was concerned about the pressure that their system might place on researchers.

The most common concern regarding the outputs of the system, mentioned by a six groups was a lack of comparability between panels, although one group suggested that

comparability was unnecessary. Five groups identified a failure of resources to reflect results as a concern, with three groups worried about grade inflation and one about the misuse of results for purposes not related to research excellence.

Hoped for Changes in UK Research

The workshop finished by closing the circle from the principles of assessment back to research culture, by asking the groups what changes in research culture they hoped their system would bring about.

The most common hope for a new research assessment system was that it would encourage and foster interdisciplinary work and allow more scope for research innovation - a theme that was mentioned by almost one third of groups. One group pointed out that the RAE's narrow subject specific assessment was at odds with the Research Councils' and Government's increasing emphasis on transdisciplinary research. It was also hoped that reducing the burden of the RAE and decreasing the scope for game playing would reduce the level of distraction from research. It was felt that submission of all researchers would improve research culture by providing the institution with an incentive to mentor and develop under performing researchers.

This incentive to develop staff was seen as part of a wider hope that the new system would lead to better nurturing of young research staff and a more inclusive research culture of equal opportunities. Some groups also hoped that their systems would lead to a research culture where a wider range of research institutions with more diverse missions could flourish.

It was hoped, by six groups, that reducing the step changes in funding caused by the RAE the new systems would improve institutions ability to plan. This was thought to be a particular advantage of the two systems where the panel had an explicit developmental role. Six groups also felt that the new system needed to be dynamic and able to react rapidly to changes in the research landscape.

Seven groups returned to the theme of transparency and consultation, suggesting that improvements in these areas would increase the sense of ownership by the academic community and lead to increased trust.

Although one group felt the RAE process was inevitably a zero sum game, other groups thought that data from a rigorous and credible assessment process could provide valuable data to increase government funding for research. There was also a feeling that an improved assessment system, with an increased credibility born out of transparency and rigorous process, could improve the international standing of UK research. The key aspects of a research assessment framework that the researchers hope would bring about these improvements in research culture are described in the executive summary at the start of the report.

Appendix A – Methodology

This appendix describes the methodology used in the workshops. It covers the methodology selection and implementation; how the participants were selected and describes workshop process and data analysis techniques.

Methodology Selection

The project required a method that could rapidly collect the views and opinions of heterogeneous groups of people. Their views needed to be collected in a semi-structured format, that would allow comparison between workshops groups, and between different groups within the participants. The facilitated workshop methodology provides a method of achieving this; and has been successfully used, with a similar spread of participants, to investigate attitudes to scientific careers[10]. The project methodology can be split into four phases that are detailed below: workshop design; participant selection; workshop process; and analysis of results.

As with all qualitative research methodologies the facilitated workshop provides a method of gaining an insight into attitudes and opinions; however, this insight is – by its very nature – not necessarily representative of the population sampled in a statistical sense. The structured nature of the workshops may have prevented the completely free expression of participants' views; however, by using a relatively flexible and free form structure we attempted to avoid this.

Workshop Design

Jonathan Grant and Steven Wooding devised the structure of the workshop. This structure was then refined after a 'dry run' in which RAND Analysts played the part of academics.

Participant Selection and Recruitment

In England, the Higher Education Funding Council for England (HEFCE) used its contacts in Higher Education Institutions, to nominate representatives to attend the workshops. HEFCE's regional representatives also organised the venue and facilities for the workshops. In Scotland, Wales and Northern Ireland the organisations responsible for recruitment and logistics were, respectively: The Scottish Higher Education Funding Council; The Higher Education Funding Council for Wales; and Department for Education and Learning (Northern Ireland).

After HEFCE had provided us with a list we contacted the participants by email to provide them with details of the workshop. We also included a questionnaire designed to allow us to classify the participants into groups for later analysis. If participants

---

[10] 'Radical Thinking, Creative Solutions' Career Issues in UK Academic Research, July 2001, The Wellcome Trust, available from http://www.wellcome.ac.uk/en/1/biosfgcdprad.html

failed to return the completed questionnaire, we asked them to fill in a paper copy when they attended the workshop.

Workshop Process

Outline descriptions of the different stages of the workshop are given in the relevant results section of the main body of the report; this section contains the organisational details.

The workshops took place in a variety of rooms, from teaching rooms to grand council rooms. Wherever possible we attempted to set rooms up in a cabaret style with groups of chairs clustered around tables; however, this was not always practical. In some venues we also had the use of breakout rooms, but on other occasions groups had to work in different corners of a large room.

Task 1

When participants arrived at the workshop they were allowed to sit wherever they choose, although we suggested that they not sit next to someone they already knew. For the first task we asked them to pair with their neighbour and discuss characteristics of high quality research and characteristics of research assessment systems.

Task 2

For the second task participants were divided into four groups; this division was done on the basis of where the participants were sitting, by dividing the room into four sectors. Each group was then provided with a marked up sheet of flipchart paper, on which to write their points and questions for feedback to entire workshop group.

Task 3

We wanted to ensure that the groups for this exercise contained a mix of people from different groups in the previous exercise. To achieve this, we worked around the room counting people off, for example if there were three groups: first person into group 1, second person into group 2, third person into group 3, fourth person into group 1, fifth person into group 2 etc. Each group, of between four and seven people, was provided with a flip chart on which to outline their research assessment system for feedback to the entire workshop group.

Task 4

For this task, one member of each group was rotated into the next group to provide an external viewpoint for re-examination of the assessment system designed in Task 3. For example, if there were three groups, one person was moved from group 1 to group 2; one person from group 2 to group 3; and one

person from group 3 to group 1. The participants moved between groups were either selected by volunteering, or were those how had presented their group's system to the plenary session at the end of Task 3.

Analysis

For all the tasks we have attempted to draw out themes that recurred across a number of workshops. For Task 1 we grouped the characteristics suggested by participants into clusters; and for Task 2 we similarly grouped the good points, bad points and questions. We carried out this clustering by printing each individual suggestion on a separate sticky label and then arranging these labels into groups – this technique allowed continual refinement and revision of clustering during the grouping process.

For Tasks 3&4 we developed the technique described above by writing the suggested modifications to the system of Expert Review, onto larger sticky labels. If another group suggested the same modification, we noted this on the first sticky label. We then arranged these sticky labels around even bigger labels that described emerging themes. Points made by groups in Task 4 where then overlaid onto the Task 3 collage, with any points that could not be placed being analysed separately. Part of the final collage is shown on the cover of this report, with the cluster of labels around 'Transparency and Openness' clearly visible in orange just above the title.